Robust Statistical Model-Based Cell Image Interpretation[1]

P. Zhou and D. Pycock
School of Electronic and Electrical Engineering
The University of Birmingham
Email: [P.Zhou, D.Pycock]@bham.ac.uk

Abstract

A robust and adaptable model-based scheme for cell image interpretation is presented that can accommodate the wide natural variation in the appearance of cells. This is achieved using multiple models and an interpretation process that permits a smooth transition between the models. Boundaries are represented using trainable statistical models that are invariant to transformations of scaling, shift, rotation and contrast; a Gaussian and a circular autoregressive model (CAR) are investigated. The interpretation process optimises the match between models and data using a Bayesian distance measure. We demonstrate how objects that vary in both shape and grey-level pattern can be reliably segmented. The results presented show that the overall performance is comparable with that of manual segmentation; the area within the automatically detected and the manually selected cell boundaries that is not common to both is less than 5% in 96% of the cases tested. The results also show that the computationally simpler Gaussian boundary model is at least as effective as the CAR model.

1 Introduction

Many researchers have sought to automate the analysis of epithelial cell images for cancer screening and other diagnostic purposes. Whilst a number of advanced prototype systems have been announced that automate cell screening, the accurate location of the cell and nuclear boundary remains a challenging task. Like most natural objects the form of epithelial cells is highly variable. Even with a single preparation and staining protocol the grey-level contrast at the cytoplasm boundary varies from 30% to less than ½%. The pattern of grey-level values within the cytoplasm or nucleus of a cell is also highly variable. The cytoplasm of a stained cell may be uniformly dark, present a graduated shading and be finely or coarsely patterned. In addition the outline of cells may vary from being compact, to elongated, to rectangular and triangular (see Figure 5). Images of epithelial cells are further complicated by cells touching and overlapping one another (see Figure 6), by the presence of other small darkly stained cells, such as leucocytes, and by the presence of bacteria.

Many cell image analysis systems seek to overcome these problems by analysing nuclei only and using preparation protocols that give high levels of contrast[5]. Others use recognition strategies that can accommodate image segmentation errors[9]. It is well recognised that most diagnostic information is contained in the interpretation of

[1]This work is supported by the EPSRC under grant No. GR/G59448.

BMVC 1995 doi:10.5244/C.9.12

the nucleus. However, overall system performance is likely to be impaired when the presence of the cytoplasm or segmentation errors are ignored.

Relatively simple edge detection and boundary tracing methods do not work well with images of epithelial cells. Methods based on Mathematical Morphology[10] capture local structure and are often more effective but are seldom effective in processing complex scenes. More sophisticated approaches to boundary interpretation using Snakes and Active Contour methods[2, 8] have been shown to be effective in locating moderately complex boundaries. However it is often necessary to adjust their control parameters carefully to avoid inappropriate interpretations. It is also often difficult to avoid the selection of non-optimal, local minima of the energy function.

To reliably segment complex images, such as those of epithelial cells, robust adaptable models and interpretation strategies are required. Here we describe an approach based on the use of robust statistical boundary models that permits many forms of boundary to be described and provides strong mechanisms to control the process of interpretation. These models have the advantage that they can be acquired by training. In this paper we describe how object boundaries are characterised in terms of simple incremental shape and grey-level properties. We show that statistical models provide a robust representation for the variability in form of epithelial cells. We describe and compare two parametric models: a Gaussian and a circular autoregressive (CAR) model. The models developed are relatively simple in concept and their computational complexity is low so that it is feasible to compare alternative boundary interpretations. The models are deformable (not rigid) and are largely invariant to changes of scale, shift, rotation and contrast. In this work segmentation is formulated as a process that optimises the probabilistic match between a multi-facet model and data using a Bayesian distance measure.

The overall interpretation strategy is summarised in Section 2 and the statistical boundary models are defined in Section 3. Boundary selection is described in Section 4 and the results of extensive tests are presented in Section 5 with concluding comments in Section 6.

2 Overview of Interpretation Strategy

The proposed scheme of interpretation follows the sequence of progressive refinements described in Figure 1. A simple adaptive threshold scheme is used for Step 1 of the algorithm. The generation of candidate boundary points in Step 2 uses a

1. Locate dark regions (the centres of potential epithelial cells).
2. Generate a table of candidate boundary points at regular intervals through 2π.
   2.1. For each radial direction generate a row of candidate boundary points.
3. For each candidate boundary point in the table compute a set of feature measures (as defined in Section 3.1).
4. Select the best model and boundary combination.
   4.1 For each boundary model select the best boundary.
      4.1.1 For each candidate on the first radial search path select a boundary
         4.1.1.1 For each remaining radial path select a boundary point.
            i) Compute a set of similarity measures (see Section 3.1).
            ii) Combine the evidence from each similarity measure and the grey-level edge strength.

Figure 1. Algorithm summary.

method similar to that described in previously reported work[1] and is summarised in Figure 2. The strategy adopted assumes that the boundary is not grossly convoluted, i.e., it is assumed that each radial path will cross the boundary once only. Steps 3 and 4 of the algorithm are described in detail in Sections 3 and 4.

1. For each radial scan direction:
   - 1.1 Sample a radial profile.
   - 1.2 Convolve the profile with the Laplacian of a 1-D Gaussian function and determine zero crossings along the profile.
   - 1.3 Convolve the profile with the 1st derivative of the Gaussian function and select the L highest magnitude responses at these zero crossings.

Figure 2. Generation of candidate boundary points.

3 Boundary Modelling

3.1 Boundary Feature Definition

A boundary is approximated by an ordered sequence of points, $B=\{b_1, b_2,..., b_i,..., b_N\}$, sampled from the boundary at regularly spaced angle about the nominal centre of the boundary, as shown in Figure 3. Since the boundary is closed $b_i = b_{i+N}$.

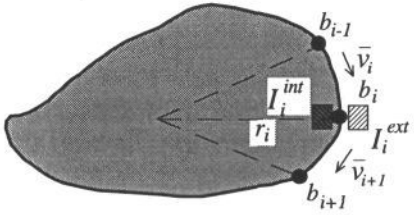


Figure 3. Definition of features for boundary representation.

To characterise the shape and grey-level properties of a boundary the following are used:

i) radial distance from the "centre" to the boundary point, r_i ;
ii) boundary curvature computed as the change in boundary direction, cv_i ;
iii) the pattern of grey-level values, $I^{int}{}_i$, in a (5×5 pixel) rectangular region inside the boundary, as described by μ_i^{int} and σ_i^{int};
iv) the pattern of grey-level values, $I^{ext}{}_i$, in a (5×5 pixel) rectangular region outside the boundary, as described by μ_i^{ext} and σ_i^{ext} ;
v) grey-level edge strength, mlr_i ;
vi) grey-level edge direction, ϕ_i .

The scale invariant curvature measure used is:

$$cv_i = \left| \frac{\bar{v}_i}{|\bar{v}_i|} - \frac{\bar{v}_{i+1}}{|\bar{v}_{i+1}|} \right|^2$$

where $\bar{v}_i$ is the vector from b_{i-1} to b_i.

From these basic measures a set of similarity measures are defined to provide invariance. Radial similarity is defined as:

$$drs_i = \frac{r_i}{r_{i-1}}$$

For clarity I_i represents $I^{int}{}_i$ and I_i^{ext} respectively and the pattern of grey-level values are modelled by the Gaussian process, $I_i \sim N(\mu_i, \sigma_i)$. To achieve invariance the similarity measures for μ_i and σ_i are defined as:

$$\mu s_i = \frac{\mu_i}{\mu_{i-1}} \quad \text{and:} \quad \sigma s_i = \frac{\sigma_i}{\sigma_{i-1}}$$

Similarity measures are also defined for edge strength using a maximum likelihood ratio estimate[11]:

$$mlr_i = \frac{\hat{\sigma}_i^M}{\hat{\sigma}_i^m \hat{\sigma}_i^{M-m}}$$

where: M, m and M-m denote the total population and the two sub-populations of a profile sample as shown in Figure 4.

σ_i^M, σ_i^m and σ_i^{M-m} represent the estimated standard deviation of the respective populations.

The resulting similarity measure for edge strength is:

$$mlrs_i = \frac{mlr_i}{mlr_{i-1}}$$

Edge direction is determined by rotating a rectangular sample region about the position of a boundary point to locate the direction of maximum edge response at that position (see Figure 4). Edge response is defined in terms of the maximum likelihood ratio[11], as above. Given this edge direction, ϕ_i, direction similarity is defined as:

$$\phi s_i = \frac{mod\, ulo_{2\pi}(\phi_i - \phi_{i-1})}{2\pi}$$

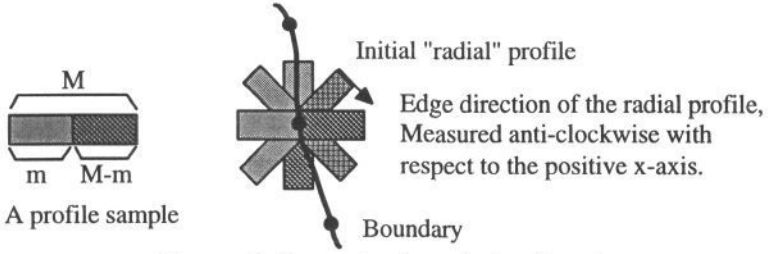


Figure 4. Determination of edge direction.

3.2 Gaussian Boundary Model

It has been observed that distributions of the above similarity measures for a class of cells are approximately Gaussian. Therefore the boundaries are modelled in terms of the mean and variance for each similarity measure.

3.3 Circular Autoregressive Boundary Model

A circular autoregressive model[2] (CAR) allows the relationship between several successive states of a stochastic sequence to be expressed. CAR procedures can be used for prediction and classification and have been extensively used for shape recognition. The formulation used here is a simplified version of that reported previously[4, 7]. This

simplification is possible due to the use of the invariant similarity measures defined in Section 3.1.

A CAR model of order M is a prediction of the current state using a linear combination of M previous states and an error term. Given the set of observations $X=\{x_{i-M}, x_{i-M+1}, ..., x_{i-1}\}$ the predicted value, x_i, is:

$$x_i = \sum_{j=1}^{M} \theta_j x_{i-j} + \sqrt{\eta}\varepsilon_i \quad i = 1,...,N$$

where ε_i is a random, uncorrelated sequence independent of x_i with zero mean,

$\theta_1,...,\theta_M$ *and* η are model parameters to be estimated.

These parameters can be estimated from observations, $X=\{x_i\}$, by minimising the expected value, η, of the sequence error, i.e., solving, $\partial\eta/\partial\theta_i = 0$, for j=1, 2, ..., M. The term η, formulated as a mean squared error estimate, is:

$$\eta = \frac{1}{N}\sum_{i}^{N}\left[x_i - \sum_{j=1}^{M}\theta_j x_{i-j}\right]^2$$

The solution of this equation is:

$$\begin{bmatrix}\theta_1\\ \vdots \\ \theta_M\end{bmatrix} = \begin{bmatrix} \sum_i^N x_{i-1}^2 & \sum_i^N x_{i-1}x_{i-2} & \cdots & \sum_i^N x_{i-1}x_{i-M} \\ \vdots & \ddots & & \vdots \\ \sum_i^N x_{i-1}x_{i-M} & \cdots & & \sum_i^N x_{i-M}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_i^N x_i x_{i-1} \\ \vdots \\ \sum_i^N x_i x_{i-M}\end{bmatrix}$$

When $\varepsilon_i \sim \mathcal{N}(0,\eta)$ then $x_i \sim \mathcal{N}\left(\left(\sum_{j=1}^{M} q_j x_{i-j}\right),\eta\right)$ and the likelihood function for the observations, X, is:

$$p_{CAR}(\{x_i,\cdots,x_N|\theta_1,\cdots\theta_M,\eta\}) = \left(\frac{1}{2\pi\eta}\right)^{\frac{N}{2}} \exp\left\{\frac{-1}{2\eta}\sum_{i=1}^{N}\left(x_i - \sum_{j=1}^{M}\theta_j x_{i-j}\right)^2\right\} \quad (1)$$

3.4 Generic Model Description

Let the similarity measures be denoted by $X^k = \{x_i^k\}$, where $k \in [1,..., K]$ is the k-th similarity measure. Then a model, M_w, for class w, is defined by $\{f(\overline{\alpha}^k)\}$ which may be either a set of Gaussian or a set of CAR pdfs.

For the Gaussian model:

$$f(\overline{\alpha}) = \mathcal{N}(\mu,\sigma)$$

where $\overline{\alpha} = [\mu,\sigma]$

and for the CAR model (defined in Equation (1)):

$$f(\overline{\alpha}) = p_{CAR}(\theta_1,...,\theta_M,\eta)$$

where $\overline{\alpha} = [\theta_1,...,\theta_M,\eta]$

4 Boundary Selection

4.1 Interpretation of Similarity Measures

A method is required to select the best boundary from the table of L edge responses generated along each of N edge search paths, represented in an $N \times L$ table of cues, denoted by $C=\{c_{ij}\}$; where c_{ij} is the j-th candidate on the i-th radial line, $i=1,...,N$ and $j=1,...,L$. The best boundary is selected on the basis of the evidence provided by edge strength, the collection of similarity measures and the constraints of a statistical model, M_w.

Let $p(c_{ij})$ be the probability that a boundary cue, c_{ij}, is a valid boundary point. We assume that the boundaries under consideration are relatively compact and that there can be only one true boundary point in each radial direction. Thus set of candidates on each radial path, $\{c_{i1}, c_{i2}, ..., c_{ij}, ..., c_{iL}\}$, is mutually exclusive and collectively exhaustive. Therefore the probability that any one boundary cue is correct, $p(c_{ij})$, is constrained by $\sum_j p(c_{ij}) = 1$.

Given edge strength, $|e_{ij}|$, the conditional probability that c_{ij} is a valid boundary point is:

$$p(c_{ij}||e_{ij}|) = \frac{|e_{ij}|}{\sum_j |e_{ij}|}$$

Let x_{ij}^k be the similarity measure for parameter k of cue, c_{ij}, on radial line, i, and $f(\bar{\alpha}^k)$ be the pdf for model M_w (Gaussian or CAR). Then the conditional probability that c_{ij} is a true boundary point is:

$$p(c_{ij}|x_{ij}^k, f(\bar{\alpha}^k)) = \frac{f(x_{ij}^k|\bar{\alpha}^k)}{\sum_j f(x_{ij}^k|\bar{\alpha}^k)}$$

Assuming that each source of evidence is independent then the nett probability that, c_{ij}, is a true boundary point is computed as:

$$p(c_{ij}) = \frac{p(c_{ij}||e_{ij}|)\prod_k p(c_{ij}|x_{ij}^k, f(\bar{\alpha}^k))}{\sum_j p(c_{ij}||e_{ij}|)\prod_k p(c_{ij}|x_{ij}^k, f(\bar{\alpha}^k))}$$

4.2 Boundary Selection

For each boundary model, M_w, the best boundary, B_l, is determined from a set of candidate boundaries. A candidate boundary, B_j, is selected from the table of candidate boundary points, $C=\{c_{ij}\}$. Taking a boundary cue on the first row of the cue table as a start point for a path (i.e., c_{ij}, where $i=1$, $j \in [1,..., L]$) the probability that each candidate on the next row is a valid boundary point (i.e., $p(c_{ij})$ where $j=1, ..., L$) is computed as in section 4.1. The candidate boundary point, $c_{i+1,t}$, with the largest value for $p(c_{i+1,t})$ is selected as the development of the boundary curve from c_{ij}. In this way boundary points on subsequent rows are selected to obtain a candidate boundary $Bj=\{c_{1j},..., c_{Nq}\}$ where $j,..., q \in [1,..., L]$. Repeating the above process for each candidate on the first row of the table results in a set of possible segmentations, $\{B_j\}$. The best segmentation, B_l, is determined as:

$$P(B_l|M_w) = \frac{1}{N}\max_j\{c_{1j}+\cdots+c_{Nq}\}$$

4.3 Selecting Best Boundary Interpretation

The best interpretation is obtained by selecting the best boundary (generated as above) for each model. A model, M_w, comprises a set of pdfs, $\{f(\overline{\alpha}^j)\}$, and the pdfs for a boundary, B_j, are estimated as $\{f(X_l^j|\hat{\overline{\alpha}}_l^j)\}$. The distance between a boundary and a model is computed, using a Bayesian metric, as:

$$D(B_l, M_w) = \frac{\sum_j \sum_i [f((x_i^j)_l|\hat{\overline{\alpha}}_l^j) - f((x_i^j)_l|\overline{\alpha}_w^j)]^2}{\sum_j \sum_i [f((x_i^j)_l|\hat{\overline{\alpha}}_l^j) + f((x_i^j)_l|\overline{\alpha}_w^j)]^2}$$

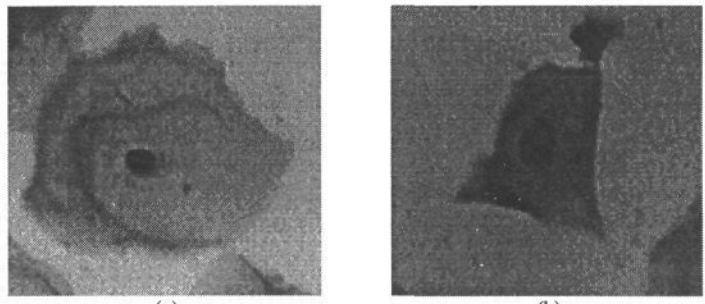
The best match, B_s, is identified by :

$$D(B_s, M_h) = \min_{l,w} [D(B_l, M_w)]$$

5 Test Results

5.1 Parameters of the Gaussian Boundary Model

The cells shown in Figure 5 were used to illustrate the behaviour of selected boundary parameters for a Gaussian model. The mean and standard deviation for each similarity measure are given in Table 1. In general: a smooth shape has a low value of mean curvature and a low standard deviation of curvature; a circular shape has a mean radial similarity value that is close to 1 and a low standard deviation of radial similarity. A uniform region of cytoplasm (and a uniform background) have a low standard deviation of grey-level similarity.



(a) (b)

Figure 5. Typical variations in the appearance of cells.

Table 1. Selected boundary parameter statistics for each of the cells shown in Figure 5.

Similarity measure		rs_i	cv_i	μs^{int}_i	σs^{int}_i	μs^{ext}_i	σs^{ext}_i
Figure 5 (a)	μ	*1.00*	*0.07*	*1.00*	*1.06*	*1.00*	*1.09*
	σ	*0.10*	*0.09*	*0.03*	*0.42*	*0.02*	*0.46*
Figure 5 (b)	μ	*1.02*	*0.08*	*1.00*	*1.11*	*1.00*	*1.61*
	σ	*0.19*	*0.18*	*0.06*	*0.49*	*0.08*	*2.82*

5.2 Cell Segmentation

Extensive tests were conducted on images of undispersed cells stained with heamatoxylin at an objective magnification of ×*60*. These images contain cells that vary greatly in shape: some are round or elongated and smooth; and others have boundaries with sharp corners. The grey-level patterns in the cytoplasm of these cells also vary considerably: some are uniformly grey, others are textured and some have a very low boundary contrast Two groups of cells are shown in Figure 6 with both the nuclei and the cytoplasm segmented. The performance of the algorithm described here was evaluated by segmenting the cytoplasm of 107 unfamiliar cells and comparing the results with those for manually segmentation.

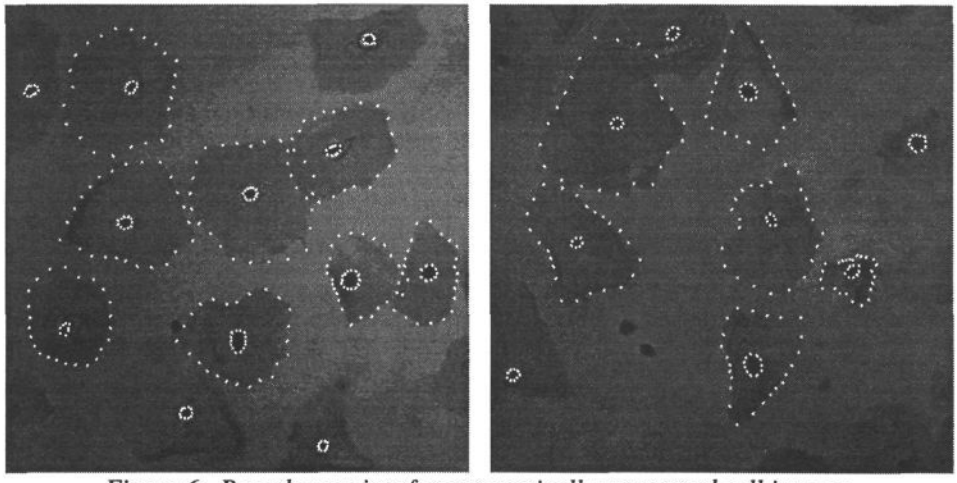


Figure 6. Boundary points for automatically segmented cell images.

5.2.1 Evaluation of Manual Segmentation

In this experiment 5 cells were manually segmented 20 times to identify the boundary between the cytoplasm and the background. In each case the average boundary was taken as the reference standard. Two "error" measures were computed. The first is the radial distance (in pixels) of each manually delimited boundary to the reference boundary, *the radial distance difference.* The second is a measure of the area within the reference boundary and each observation of the manual boundary that was not common to both. The measure used was the ratio of the area that is not common to both boundaries with respect to the reference boundary, *the area difference ratio.* The mean and standard deviation for each "error" measure are shown in Table 2.

Table 2 Consistency of Manual Segmentation.

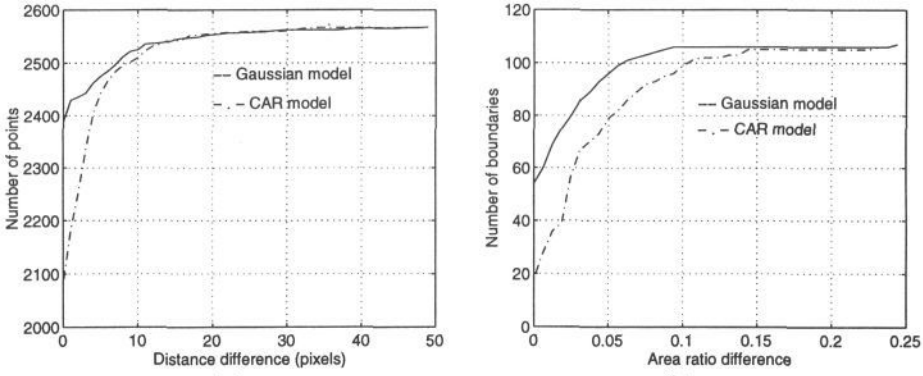
		simple isolated cells (1 cell 20 times)	low contrast, isolated cells (2 cells 20 times)	overlapping cells (2 cells 20 times)
Radial distance	μ_d	0.905	1.840	3.080
	σ_d	0.704	1.160	2.130
Excluded area	μ_R	0.026	0.036	0.055
	σ_R	0.006	0.009	0.020

Assuming that the errors follow a Gaussian distribution the data in Table 2 shows that the radial variation for manual segmentation is between 1 and 2 pixels 66% of the time and between 4 and 10 pixels 99% of the time. It is interesting to note that for manual segmentation it is significantly easier (by a factor of 3) to correctly identify the boundary of isolated high contrast cells than the boundary of overlapping cells.

5.2.2 Evaluation of Automated Segmentation

In this evaluation the cytoplasm of 107 cells in 20 images taken at random from one slide were each segmented automatically and manually. In these images there were 27 overlapping cells and 80 isolated cells (most of which were of low contrast).

Each boundary was represented by approximately 24 points and 16 models were used. The difference between each manual and automatically segmented boundary was computed to give the results summarised in Figure 7. The measures used are *radial distance difference* and *area difference ratio.* These plots show that there is little difference in the results obtained using Gaussian and CAR models.



(a) *(b)*

Figure 7. A comparison of manual and automated segmentation using a Bayesian distance metric for both Gaussian and second order CAR boundary models. a) the distribution of radial distance difference and b) The distribution of area ratio difference.

From the figures given in Table 1 we can observe that a reasonable expectation for manual segmentation is a *radial distance difference* of below 4 pixels and *area difference ratios* of below 0.05. Considering the results for automated segmentation compared to manual segmentation given in Figure 7 there are 114 boundary points with radial distance differences of more than 3.5 pixels, i.e. 4% and 5 boundaries with an area difference of greater than 0.05, i.e. 5%. This implies that the error in automated segmentation, affects, on average, 4% of the boundary of 5% of cells. Thus the precision of automated segmentation is shown to compare favourably with manual segmentation. In absolute terms Figure 7 also shows that there are very few cases of large errors in the automated segmentation procedure. Given the number of touching and overlapping cells included in the test set this indicates a high level of reliability for the automated segmentation procedure. However it is clear that the overall performance of the automated segmentation procedure is not perfect.

6 Conclusion

Gaussian models with a Bayesian distance metric have been shown to be an effective and robust adaptive model for describing cell boundaries. The Gaussian model and Bayesian distance metric is to be preferred as the computational cost of using this representation is less than that for the CAR model. The computational complexity of the boundary selection process was minimised using dynamic programming. It was found to take approximately 3 seconds to segment the cytoplasm of one cell on a SUN SPARCStation 10 with a single 40 MHz processor.

In the work described here the angular location of boundary points is fixed. This does not appear to cause a serious problem. An alternative strategy in which the density of boundary points is variable is described elsewhere [11].

Further work is required to compare the performance of this algorithm with others for cell image segmentation, to evaluate performance on other biological images and to investigate ways of accelerating the process. Work is already in hand on high-level reasoning strategies to guide interpretation.

7 References

[1] Azzopardi PJ, Pycock D, Taylor CJ and Wareham AC, "An Experiment in Model-Based Boundary Detection", in *Proc. AVC88*, Manchester, 1988, pp31-36.

[2] Cohen LD and Cohen I, "Finite-Element Methods for Active Contour Models and Balloons for 2-D and 3-D Images", *IEEE Trans. Patt. Anal. Machine Intell.*, **PAMI-15**, 11, 1993, pp 1131-1147.

[3] Davis MHA, Vinter RB, *Stochastic Modelling and Control*, Chapman and Hall, London, 1985, pp .

[4] Dubois SR, Glanz FH, "An Autoregressive Model Approach to Two-dimensional Shape Classification", *IEEE Trans. Patt. Anal. Machine Intell.*, **PAMI-8**, 1, 1986, pp 55-66.

[5] van Driel-Kulker AMJ and Ploem-Zaaijer JJ, "Image cytometry in automated cervical screening", *Anal. Cell. Path.*, **1**, 1, 1989, pp63-77.

[6] Duda RO and Hart PE, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, 19783, pp 44-49.

[7] Kashyap RL, Chellappa R, "Stochastic Models for Closed Boundary Analysis: Representation and Reconstruction", *IEEE Trans. Inform. Theory,* **IT-27**, 5, 1981, pp 627-37.

[8] Kass M, Witkin A, Terzopoulos D, "Snake: Active Contour Models," *International Journal of Computer Vision*, **1**, 4, 1988, pp 321-31.

[9] Lee JSJ, Bannister, MS, Kuan LC, Bartels PH and Nelson AC, "A Processing Strategy for Automated Papanicolaou Smear Screening", *Anal. Quant. Cytol. and Histol.*, **14**, 5, 1992, pp415-425.

[10] Meyer F, "Automatic Screening of Cytological Specimens", *CVGIP*, **35**, 1986, pp356-369.

[11] P Zhou and Pycock D "Robust Model-Based Boundary Cue Generation for Cell Image Interpretation", *Proc. BMVC'95*, Birmingham 1995.